What makes a face photo a 'good likeness'?

Kay L. Ritchie[1,2], Robin S. S. Kramer[1,2] & A. Mike Burton[1]

[1] Department of Psychology, University of York, UK

[2] School of Psychology, University of Lincoln, UK

Correspondence to:
A. Mike Burton,
Department of Psychology,
University of York,
York,
YO10 5DD, United Kingdom.
mike.burton@york.ac.uk

Abstract

Photographs of people are commonly said to be ‘good likenesses’ or ‘poor likenesses’, and this is a concept that we readily understand. Despite this, there has been no systematic investigation of what makes an image a good likeness, or of which cognitive processes are involved in making such a judgement. In three experiments, we investigate likeness judgements for different types of images: natural images of film stars (Experiment 1), images of film stars from specific films (Experiment 2), and iconic images and face averages (Experiment 3). In all three experiments, participants rated images for likeness and completed speeded name verification tasks. We consistently show that participants are faster to identify images which they have previously rated as a good likeness compared to a poor likeness. We also consistently show that the more familiar we are with someone, the higher likeness rating we give to *all* images of them. A key finding is that our perception of likeness is idiosyncratic (Experiments 1 and 2), and can be tied to our specific experience of each individual (Experiment 2). We argue that likeness judgements require a comparison between the stimulus and our own representation of the person, and that this representation differs according to our prior experience with that individual. This has theoretical implications for our understanding of how we represent familiar people, and practical implications for how we go about selecting images for identity purposes such as photo-ID.

Keywords: Facial likeness; familiarity; iconic images; face averages; prototype; mental representation.

1. Introduction

We all understand what it is to say that a particular image of someone is a good likeness. In fact, likeness is considered important for official forms of photo-ID, with passport-issuing offices around the world requiring someone familiar with the applicant to verify the likeness of the passport image (Australian Passport Office, 2012; Passport Canada, 2013; HM Passport Office, 2014). Despite this, there has been no systematic investigation of why observers pick out one image as a good likeness while considering another to be a bad likeness.

A number of different techniques in research on face recognition have used the concept of likeness as a key measure in the success of their manipulation. For example, studies manipulating the distinctiveness of face images (Allen, Brady & Tredoux, 2009; Lee & Perrett, 2000), research on the caricaturing effect (Benson & Perrett, 1991; Rhodes, Brennan & Carey, 1987), and research on face composites (Bruce, Ness, Hancock, Newman & Rarity, 2002; Frowd et al., 2014) all used ratings of likeness as the dependent measure. Yet none have defined what this term means, relying on the fact that we all, participants and readers alike, intuitively understand the concept of likeness. Such an understanding seems to rely on the notion that a good likeness closely matches a canonical representation of a known person, or is perhaps some kind of 'super-stimulus' providing efficient access to such a representation, as has sometimes been claimed for caricatures.

When we consider different types of photos, we can see that one person can look very different across images depending on what they are trying to achieve in each (e.g., Leikas, Verkasalo, & Lönnqvist, 2013). For example, someone's passport photo will look different from their work website photo or their holiday photos. Previous work has shown that observers familiar with the person pictured can easily see that multiple, varied photos show the same person, whereas unfamiliar observers struggle to identify the same person across multiple images (e.g., Jenkins, White, van Montfort & Burton, 2011; Laurence & Mondloch, 2016). The same is true even when only two images are pictured side by side and observers are asked to indicate whether the images show the same person or two different people. This task is easy for familiar viewers but significantly more difficult for unfamiliar viewers (Bruce, Henderson, Newman & Burton, 2001; Clutterbuck & Johnston, 2002, 2004; Ritchie et al., 2015). Here, the difference between familiar and unfamiliar viewers seems to be their

capacity to cope with a range of variability across different photos of the same person. With increased familiarity comes an ability to recognise a person from an increased range of images. Here we test whether familiarity also leads to an increased tolerance to the range of images one would categorise as being a good likeness.

In an earlier systematic exploration of our ability to recognise and provide information (such as occupation) about familiar people from face images, the most common reason participants gave for failing to recognise a familiar person (71% of recognition failures) was that the photo was a 'bad likeness' (Hay, Young & Ellis, 1991). The authors go on to quote participants' responses, giving examples such as "that's not what I remember him looking like", and "she's much younger in that photograph" (Hay, Young & Ellis, 1991, p. 778). One of the most interesting observations made was that different participants gave bad likeness responses to different images. This suggests that there is not something inherently poor about any given image, leading to all participants failing to recognise the same images. Rather, when each individual tries to recognise a familiar person, they compare the image under consideration with their own representation of that person. Each individual's representation of a person differs due to different levels of exposure to, or familiarity with, each target person, and so different images are a poor likeness for different observers.

Perhaps the most familiar face for a person is their own, and a recent study has shown that the images which participants select as a good likeness of themselves are, in fact, not optimal for identifying them (White, Burton & Kemp, 2016). Participants ranked images of their own face for likeness, and a separate group of participants who were previously unfamiliar with these people also ranked the images for likeness after seeing a short video clip of each person. The images selected by this group as a good likeness, along with those selected by the participants themselves, were then used in a face matching task where two images were presented side by side, and a new group of participants were asked to judge whether or not the two images showed the same person. The images chosen by the participants themselves yielded lower accuracy on the matching task than the images selected by (unfamiliar) others. The results show that likeness judgements change as familiarity with the person pictured changes.

Jenkins et al. (2011) investigated the relationship between familiarity with celebrities and likeness ratings for multiple images of those celebrities. Participants rated multiple images of

celebrities for likeness on a seven-point scale (extremely poor to extremely good likeness). Mean likeness ratings for images of each celebrity were positively correlated with the percentage of participants who recognised each celebrity. The results also showed that the variability in likeness ratings for multiple images of one celebrity could be greater than the variability in ratings between celebrities. Importantly, this approach used group means across participants. However, following from the earlier finding by Hay et al. (1991) that participants rated different images as a poor likeness, it may be more beneficial to analyse likeness ratings for each participant individually, rather than using group means.

What does it mean, therefore, for an image to be 'a good likeness'? Likeness judgements rely on a comparison between a physical stimulus and our own representation of an identity. The nature of this representation is unclear. As we become familiar with someone, we see multiple, variable instances of them which we need to be able to reconcile as the same person. Previous research has shown that when presented with multiple images of a person, viewers automatically extract the mean of the set (Kramer, Ritchie & Burton, 2015), and it has been suggested that these averages provide us with stable representations of people (Burton, Jenkins, Hancock & White, 2005; Frowd et al., 2014; Robertson, Kramer & Burton, 2015). We might also represent some familiar celebrities through an 'iconic' or famous image (Allen et al., 2009; Carbon, 2008). In an experiment using common and uncommon images of celebrity faces, participants were able to recognise and name celebrities from commonly-seen or iconic images (~80% accuracy), but recognition rate dropped dramatically (~25%) when participants were shown uncommon images of the celebrities (Carbon, 2008). The results suggest that we recognise some celebrities from specific, commonly-seen images.

In the current research, we have sought to explore factors underlying likeness judgements for familiar faces. Underpinning this work is the idea that an image which is rated as being a good likeness should more closely resemble the rater's idea of what that person truly looks like in comparison with an image which is rated as a poorer likeness. We build on previous research which has shown that observers recognise a familiar individual across many variable images, suggesting that each of these images resembles their idea of what this person truly looks like. This study also follows on from the finding of Jenkins et al. (2011) that the more participants in a group who were familiar with a given celebrity, the higher the group likeness ratings for multiple images of that celebrity. These premises lead to three testable predictions: 1) it will be easier for an observer to recognise someone from an image which they perceive

to be a good likeness of that person; 2) raters who are highly familiar with a person will have seen many images of them, and so will give higher likeness ratings to a larger range of images than will a less familiar viewer; 3) the specific images that each rater gives high likeness ratings to will be linked to their own experience of each person, and so likeness ratings will be different from one observer to the next.

In Experiment 1, we addressed the relationship between familiarity with a celebrity and likeness ratings given to images of them, as well as the idiosyncratic nature of likeness ratings. We examined individual and group variance in ratings (Hönekopp, 2006) to investigate the idiosyncratic nature of the perception of likeness. In addition, participants completed a speeded name verification task. If the perception of likeness from images maps on to the mental representation of identities, then it is reasonable to hypothesise that the higher the likeness rating, the faster that image will be verified as picturing the named identity. Experiment 2 explored the association between likeness ratings and observers' own prior experiences with each celebrity using images from films which participants had or had not seen. Finally, Experiment 3 looked at other types of images which have previously been suggested as candidates for representing familiar people (face averages) and certain celebrities (iconic images), and tested whether these types of images are given higher likeness ratings than other images.

2. Experiment 1 – The idiosyncratic nature of likeness perception

In order to test whether the perception of likeness is specific to each individual observer, we used a technique which allows us to differentiate between individual and group variance in likeness ratings. We asked observers to rate images for likeness twice, allowing us to determine whether variance in likeness ratings is explained predominantly by private or shared variance (following Hönekopp, 2006). We hypothesised that private perception of likeness (each person's idiosyncratic perceptions) would explain more variance than shared perceptions of likeness (agreement across the group) since what constitutes a good likeness could be different for each observer. If the perception of likeness is tied to each observer's familiarity with the person in question, observers may give higher likeness ratings for images of highly familiar compared with less familiar celebrities. We also tested whether likeness ratings were related to ease of recognition using a speeded name verification task. After rating the images for likeness, participates were shown the name of a celebrity followed by

either an image of them or of another celebrity. If images rated as a good likeness are more easily recognised, we predicted that reaction times on the speeded name verification task would be faster for those images previously rated as a good compared to a poor likeness.

2.1. Method

2.1.1. Participants

Thirty-five participants (5 men; mean age: 19 years, range: 18-32 years) took part. All were students or other members of the University of York, UK, or the University of Lincoln, UK.

2.1.2. Materials and Procedure

We used 15 images of each of 5 male Hollywood celebrities (Brad Pitt, Hugh Jackman, Matt Damon, Tom Cruise, and Tom Hanks). Hollywood celebrities were chosen to maximise the likelihood that all participants were familiar with all celebrities. Images were taken from a Google Image search specifying high resolution, broadly front-on face images. These varied in environmental factors such as lighting, head angle, expression, as well as person-based factors such as age, facial expression, hairstyle, etc. (see Figure 1 for examples). All images were cropped to show just the head, although backgrounds were not removed. Participants began with a familiarity measure in which they were shown the name of each celebrity and asked to rate how familiar they were with each person on a scale from 1 (not very familiar) to 7 (very familiar). In the ratings task, participants viewed all 75 images in a random order, and were asked to indicate how good a likeness of that celebrity each image was. The likeness rating used a 1 to 7 scale from ‘very bad likeness’ to ‘very good likeness’. After rating each image once, participants had a short break before rating each image a second time (again, in a novel random order). Participants also rated the images for trustworthiness and dominance, the results of which are not presented here.

Figure 1. Example stimuli for Experiment 1. Multiple images showing the same identity, but varying in lighting, colouring, hairstyle, etc. [Copyright restrictions prevent publication of the actual images used, though these are available from the authors. Images in Figure 1 are illustrative of the experimental stimuli, and depict someone who did not appear in the experiments but has given permission for the images to be reproduced here.]

Following the ratings task, participants completed a speeded name verification task. A name was displayed on the screen for 1500 ms, and then replaced by an image which remained on the screen until response. On half of the trials, the face matched the name. Participants responded via button press, indicating whether the image showed the same person as the name or not, and were instructed to respond as quickly and as accurately as possible. There were 15 match trials per identity, each showing the person named, and the images previously rated for likeness. The 15 mismatch trials per name showed 5 images of three of the other identities, counterbalanced across identities to ensure each name and each person were seen

equally often. The images used for the mismatch trials were novel images which had not been seen during the likeness ratings phase of the experiment.

2.2. Results and Discussion

All participants reported knowing who each celebrity was (although their familiarity with each identity varied) prior to the experiment. For all correlations presented here, *r* values were converted to *z* using Fisher's *r*-to-*z* transformation (which corrects the skew in the distribution of *r*). These *z*-scores were then compared to zero using a two-tailed, one-sample *t*-test, where a value significantly different from zero indicates an association between the two variables at the individual level (e.g., Back et al., 2010; Kramer, Gottwald, Dixon, & Ward, 2012). Where no relationship is present, the distribution of correlations (*z*-scores) will not differ from zero. In other words, we focused on individual-level correlations and whether their distribution differed from zero, whereas researchers often average across participants and then correlate rated familiarity and likeness ratings, for example. The 'average of correlations' approach used here allows us to focus on the responses of single observers, crucial for analyses where we expect each person to respond differently based on their personal experiences, whereas individual-level mechanisms can be missed or relationships can be inflated if we average across participants and then correlate these results (e.g., Monin & Oppenheimer, 2005).

We began by correlating rated familiarity with mean likeness ratings. For each participant, we calculated a Spearman rank correlation between their rated familiarity with each celebrity and their mean likeness rating of the 15 images of that celebrity in both the first and second rating blocks. There was a general positive correlation between familiarity and mean likeness score for both participants' first ratings ($z = .230$, $t(31) = 2.243$, $p = .032$, $d = .40$ (comparing to zero), and second ratings ($z = .290$, $t(31) = 2.158$, $p = .039$, $d = .38$). These results extend those of Jenkins et al. (2011) who showed that the more participants in the group who reported being familiar with a celebrity, the higher the group likeness rating for that celebrity. Here we have shown this at the individual level – the more familiar a person is with a celebrity, the higher likeness rating they give to images of that person.

Using the two likeness ratings of each image per participant, we carried out an analysis of private versus shared variance following methods previously used for attractiveness ratings (Hönekopp, 2006). Using this analysis, we found a greater contribution of private in

comparison with shared variance when explaining participants' likeness ratings (beholder index, bi_1 = .64). Here, we considered absolute differences in judges' scores to be unimportant in that we assumed general differences between judges reflected a difference in scale use. (Note that if we considered judge-score differences to be meaningful then our indices reported here would be underestimates of the amount of private taste.) Our result means that 64% of the meaningful variance stable across time was due to a private perception of likeness, with the remaining 36% of variance shared across participants. As such, the perception of likeness is to a greater extent specific to each individual observer rather than being agreed upon across observers.

We explored the effects of familiarity and likeness ratings on the speeded naming of each celebrity image. Error rates for the speeded name verification task were very low (*mean of* 3.5 errors from 75 trials per participant). For the RT analysis, we took RTs only for name/face match pairs in the speeded name verification task to which participants responded correctly. We excluded trials for which the RT was 2 SDs above or below that participant's mean RT (M = 4.58 trials excluded per participant). For each participant, we correlated their familiarity ratings for the five celebrities with their mean reaction time per celebrity in the speeded name verification task. We then converted these Spearman r values to z and averaged across the five z values to obtain one familiarity/RT correlation value for each participant. There was a negative correlation between familiarity and mean reaction time which was significantly different from zero (mean z = -.423, $t(31) = 3.125$, p = .004, d = .552). This shows that participants are faster to identify celebrities that they are more familiar with. The images for name/face match trials in the speeded name verification task were the same images as had previously been rated for likeness, allowing us to look at the relationship between likeness ratings and reaction times for each image. For each participant, we correlated their RT for each image with their likeness rating for that image, separately for each celebrity. This gave us five correlation values for each participant. Again, by converting these to z scores, we could then take the mean of these values, obtaining one likeness/RT correlation value for each participant. The negative correlation between mean likeness rating and mean RT was significantly different from zero (mean z = -.110, $t(34) = 3.718$, p = .001, d = .628). In this experiment, we found faster reaction times for individual images which had previously been rated as a good likeness.

Here, we have shown evidence of the idiosyncrasy in judgements of likeness. Sixty-four percent of the variance in likeness ratings was explained by a private rather than shared perception of likeness, supporting our hypothesis that what constitutes a good likeness is different for each observer. These results also show that participants were faster to identify celebrities with whom they were most familiar. Furthermore, we have shown that participants' reaction times for each image reflect their likeness rating for that image, with images rated as a better likeness being identified more quickly.

To pursue the idea that the perception of likeness is specific to the observer's previous experience, we conducted a second experiment using images from films. It is possible that participants had not previously seen the images used in the first experiment presented here since the images were gathered from internet search, and typically showed the celebrities at film premieres and other public events. Although these types of images are likely to appear in magazines, the more common way of seeing these Hollywood celebrities is starring in films, portraying particular characters. We carried out a second experiment using images from films in order to tie the experimental images to participants' own previous experience of the celebrities through the films they have seen.

3. Experiment 2 – Likeness tied to specific exposures

This experiment followed the methods of Experiment 1 but used images from specific films as opposed to images of the celebrities out of character. Our previous experiments suggest that the perception of likeness differs from one observer to the next, presumably based on their prior experience of each celebrity. Using images from films allowed us explicitly to establish whether each participant had previously seen each image, based on whether or not they had seen each film.

3.1. Method

3.1.1 Participants

Thirty-two participants (3 men; mean age: 20 years, range: 17-30 years) took part. All were students of the University of York, UK, or the University of Lincoln, UK.

3.1.2 Materials and Procedure

This experiment used images of the same five male Hollywood celebrities used in Experiment 1. Here, the images were cropped still frames from specific films. The images were obtained both from a Google Image search specifying each actor and each film, and by taking stills from scenes from the films themselves. This allowed us to establish which specific images participants had previously seen, based on which films they had watched. For each of the five celebrities, we took three images from each of five films, resulting in 15 images per identity. Participants were shown the name of each celebrity and began by indicating familiarity with each celebrity on a 1-7 scale. Participants then viewed all 75 images (15 per celebrity) in a random order and rated each image for likeness on a 1-7 scale ('very bad likeness' to 'very good likeness'). Participants rated each image once.

Following the ratings task, participants completed a speeded name verification task. As in Experiment 1, the images for the name/face match trials in the speeded name verification task were those which had been previously rated. The images used for the mismatch trials were five images of three of the other identities, and as in Experiment 1, these were novel images which had not been seen during the likeness ratings phase of the experiment. Participants were then shown the names of each of the 25 films from which the stimuli for the experiment were taken. Participants were asked to indicate which films they had seen in their entirety, i.e., not simply the trailer or some images from the film. It is possible that participants who had not seen a specific film may still have been exposed to images from that film in trailers or promotional images. Participants who had seen a given film, however, had definitely been exposed to our experimental stimuli before, and would have had more experience of each of the actors portraying each of the roles than people who had not seen the film. Therefore, by asking participants which films they had seen, we attempted to account for actual prior experience.

3.2 Results and Discussion

All participants reported being familiar with all five identities in this experiment, although, as before, the amount of familiarity varied. As in the previous experiment, we began by correlating familiarity with mean likeness ratings using an 'average of correlations' approach in order to understand mechanisms at the individual level. There was a positive correlation between familiarity and mean likeness score (mean z = .371, $t(29)$ = 3.741, p = .001,

d = .683). As in Experiment 1 above, the more familiar participants were with the celebrities, the higher the mean likeness rating they gave to images of those celebrities.

Participants reported having seen a mean of 9 out of 25 films (range 3-19). In order to establish whether our film image set related to familiarity, for each participant we recorded the number of films they reported having seen for each celebrity, and correlated this with their familiarity rating for that celebrity. This gave us five correlation values for each participant, which we converted to *z* scores and averaged together, resulting in one familiarity/films-seen correlation value for each participant. We compared these values to zero using a one-sample *t*-test. We found a significant positive correlation between the number of films seen for each identity and self-reported familiarity with that identity (mean *z*-score = .288, $t(28) = 2.157$, $p = .040$, $d = .400$). We also took the mean likeness rating given to images from films each participant had seen, and the mean likeness rating for images from films they had not seen. A paired-samples *t*-test showed a significant difference between likeness ratings for images from films participants had seen (*mean* = 5.27) and images from films participants had not seen (*mean* = 4.80), $t(31) = 4.534$, $p < .001$, $d = .814$. This indicates that the film images used in this experiment reflect participants' actual exposure with each celebrity. The more familiar they were with a celebrity, the more of our chosen films they had seen, and they rated images from films they had seen as a better likeness than images from films they had not seen.

Error rates for the speeded name verification task in this experiment were low (mean of 6.56 errors from 75 trials per participant). We excluded trials for which the RT was 2 SDs above or below that participant's mean RT ($M = 3.62$ trials excluded per participant). Again, for RT analyses, we took RTs only for name/face match pairs in the speeded name verification task to which participants responded correctly. As in Experiment 1, for each participant we correlated their familiarity ratings for the five celebrities with their mean reaction time per celebrity in the speeded name verification task. We then converted these Spearman *r* values to *z* and averaged across the five *z* values to obtain one familiarity/RT correlation value for each participant. As in Experiment 1 above, there was a significantly negative correlation between familiarity and mean RT (mean $z = -.416$, $t(31) = 4.194$, $p < .001$, $d = .741$). This indicates that participants were faster to identify the celebrities with whom they were more familiar. For each participant, we also correlated their RT for each image with their likeness rating for that image, separately for each celebrity. This gave us five correlation values for

each participant. Again, by converting these to *z* scores, we could then take the mean of these values, obtaining one likeness/RT correlation value for each participant. Again, as in Experiment 1 above, there was a significant negative correlation between mean likeness rating and mean RT (mean z = -.087) $t(31) = 3.427$, $p = .002$, $d = .606$). We find faster reaction times for individual images which have previously been rated as a good likeness.

In this experiment, we used images of celebrities in particular films, allowing us to accurately establish which specific images participants had already been exposed to. Images from films that participants had seen received higher likeness ratings than images from films participants had not seen, and the number of films seen was positively correlated with familiarity. This shows that the perception of likeness is directly related to exposure. Here, we have shown that images of each celebrity which received higher likeness ratings were also responded to faster in a speeded name verification task. This is the first clear evidence that the concept of likeness, expressed through ratings, corresponds to the observer's own representation of what a person looks like.

What we have not addressed in these first two experiments, however, is the nature of participants' own representations of each celebrity. We have argued that our likeness rating and speeded name verification tasks show higher ratings and faster RTs for images which more closely resemble each participant's own idea of what the celebrity looks like. One previous paper has suggested that we may recognise some familiar identities from specific iconic images, making these images a candidate for our mental representation of these familiar people (Carbon, 2008). Other research has suggested that face averages provide stable representations of people, and so may be a candidate for our representations of familiar people (Burton et al., 2005; Jenkins & Burton, 2011). In a final experiment, we compare these two candidate representations, and ask whether face averages or iconic images receive higher likeness ratings than other exemplars.

4. Experiment 3 – Iconic images and averages

In this final experiment, we sought to establish whether iconic images or average images are judged to be a good likeness of celebrities. We selected three celebrities for whom an iconic face image existed, and three who did not have an iconic image. We also made an average image of each celebrity by averaging together images of that person. Participants first rated

each image for likeness, and then completed a speeded name verification task for each image of each identity.

4.1 Method

4.1.1 Participants

Fifteen participants (3 men; mean age 27 years, rage: 20-39 years) from the University of Western Australia took part in the image selection phase of the experiment. Forty-two participants (17 men; mean age: 27 years, range: 18-48 years) took part in the main experiment. These were students or other members of the University of York, UK, and the University of Lincoln, UK.

4.1.2 Materials and Procedure

In the image selection phase of this experiment, we used ten images of each of 16 celebrities. Eight of these celebrities were selected by the experimenters as potentially having iconic images which best represented that person (Che Guevara, Albert Einstein, Adolf Hitler, John F Kennedy, Marilyn Monroe, Audrey Hepburn, Britney Spears, and Twiggy). Eight other celebrities were chosen as probably not having a single iconic image which best represented them (Harry Styles, Justin Bieber, Kanye West, Barack Obama, Adele, Hilary Clinton, Taylor Swift, and Jennifer Lopez). Participants were presented with an array of ten images of each identity, one identity at a time. The images were numbered onscreen from 1 to 10, and participants were instructed to record via button press which was the most iconic image of each individual, with the instructions reading “For each of the following people, please pick out the image that you think is the most iconic. That is, the image that you feel most epitomises them”. The name of each celebrity was presented above their particular image array. If participants did not recognise a celebrity, they were instructed to proceed to the next celebrity without making a response.

Identities were deemed to have an iconic image if 10 or more of the 15 participants selected the same image as iconic. Identities were classified as not having an iconic image if 5 participants or fewer selected the same image as iconic. For the three identities we selected as having iconic images (Che Guevara, Albert Einstein, Marilyn Monroe), a mean of 11.0 participants chose the same image as iconic. This was then selected as the iconic image of this identity for the main part of the experiment. The mean number of different images

selected as iconic across participants for each identity was 2.67. For the three identities we selected as not having iconic images (Harry Styles, Justin Bieber, Kanye West), a mean of 4.0 participants chose the same image as iconic. For these non-iconic identities, the mean number of images chosen as iconic across participants for each celebrity was 7.0.

The images used in the main experiment were the ten images of each of the three iconic and three non-iconic celebrities (iconic referring to those who were selected in the initial phase of the experiment as having an iconic image, non-iconic referring to those for whom there was not an iconic image). We also created a computer-generated average of the ten images for each identity separately using Psychomorph software (Tidemann, Burt & Perrett, 2001).

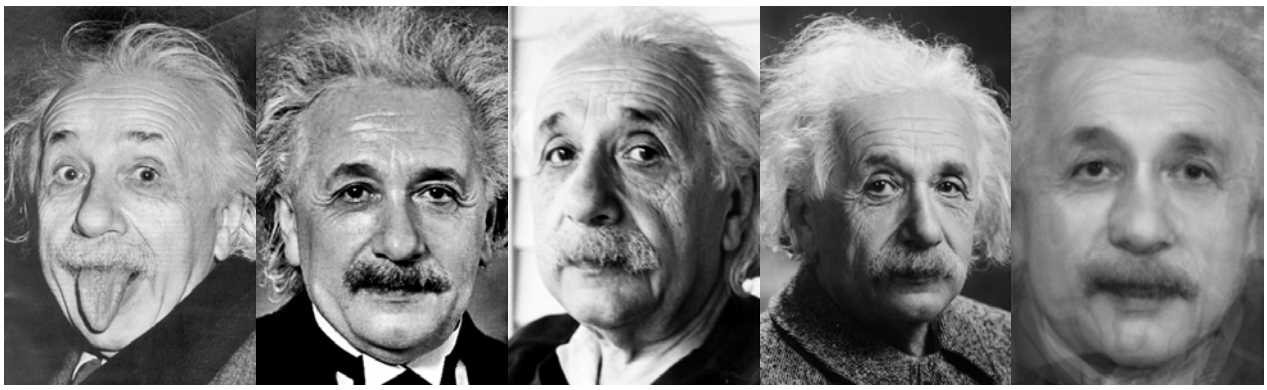

Figure 2. Stimuli used in Experiment 1. Left to right – iconic image of Albert Einstein, three of the further nine exemplars used of Einstein, and the average of all 10 exemplars of Einstein. [Images are labelled for reuse under creative commons licensing.]

For the ratings part of the experiment, participants viewed all 66 images (ten exemplars and one average image per identity). The images were presented in a random order, with each trial presented along with the instruction "How good a likeness of X is this?", where X is the name of the identity being shown. As above, participants rated each image from 1 (very low) to 7 (very high). After a short break, participants completed the speeded naming part of the experiment. There were 11 match trials per identity, each showing the person named (the exemplars and average image previously rated for likeness). There were 10 mismatch trials per name, which showed a foil celebrity chosen to match the sex, ethnicity, and approximate age of the named celebrity (foil IDs: Che Guevara–John F Kennedy, Albert Einstein–Anthony Hopkins, Marilyn Monroe–Gwyneth Paltrow, Harry Styles–Matt Damon, Justin

Bieber–Brad Pitt, Kanye West–Will Smith). The images were presented centrally, measuring 5.2 cm x 7.8 cm onscreen.

4.2 Results and Discussion

We analysed the likeness ratings separately for the iconic and non-iconic identities. For iconic identities (those identities for whom an iconic image was established in the initial image selection experiment), we took the mean likeness rating for the average image, the iconic image (as agreed upon by the 15 participants in the initial image selection experiment), and the nine other exemplars across participants. Table 1 contains the mean likeness ratings for the different image types. Data for identities with (iconic IDs) and without iconic images (non-iconic IDs) were analysed separately.

Table 1. Mean likeness ratings for different image types

Image Type	Iconic IDs	Non-Iconic IDs
Iconic Images	6.46 [6.27, 6.66]	-
Exemplars	5.11 [4.91, 5.32]	5.60 [5.38, 5.82]
Averages	3.87 [3.49, 4.23]	4.23 [3.89, 4.56]

Note: Iconic and non-iconic IDs are those who do and do not, respectively, have an iconic image, as agreed by the initial participant sample. 95% CIs given in square brackets

A one-way ANOVA for likeness ratings of iconic identities showed a significant difference in ratings between the three image types (average, iconic image, exemplars), $F(2,82) = 152.45$, $p < .001$, $\eta_p^2 = .79$, with follow-up Tukey HSD comparisons showing significantly higher likeness ratings for iconic images than exemplars and averages, and higher likeness ratings for exemplars than averages (all ps $< .05$). For identities without iconic images, a related means *t*-test showed significantly higher likeness ratings for exemplars than averages, $t(41) = 8.71$, $p < .001$, $d = 1.34$. These results show that for identities for whom there is an iconic image, that image is rated as a better likeness than other images or the average image. For identities for whom there is not an iconic image, exemplars receive higher likeness ratings than average images.

Next, we investigated the relationship between ratings and reaction times (RT) on the speeded name verification of each celebrity image. Error rates for this task were very low

(*mean of* 4.19 errors from 126 trials per participant). We excluded trials for which the RT was 2SDs above or below that participant's mean RT (*M* = 4.38 trials excluded per participant). For the RT analysis, we took RTs only for name/face match pairs in the speeded name verification task to which participants responded correctly. For each participant, we correlated their RTs for each image with their likeness rating for that image, separately for each celebrity. This gave us six correlation values for each participant, one for each celebrity. We then converted these Spearman *r* values to *z* and averaged across the six *z* values to obtain one likeness/RT correlation value for each participant. These values were then compared to zero using a one-sample *t*-test. There was a significant negative correlation between mean likeness rating and mean RT (mean *z*-score = -.120, $t(41) = 4.53$, $p < .001$, $d = 0.699$). This shows that participants were faster to verify images which they had previously rated as a good likeness.

To examine RTs for speeded name verification further, we calculated mean RTs for the different image types across participants. Reaction times on the speeded name verification task for each image type for both iconic and non-iconic identities are shown in Table 2.

Table 2. Mean RTs (ms) for different image types

Image Type	Iconic IDs	Non-Iconic IDs
Iconic Images	537 [494, 559]	-
Exemplars	570 [541, 598]	574 [540, 607]
Averages	602 [551, 651]	616 [576,567]

Note: Iconic and non-iconic IDs are those who do and do not, respectively, have an iconic image, as agreed by the initial participant sample. 95% CIs given in square brackets.

A one-way ANOVA for identities with iconic images showed a significant difference in RTs between the three image types, $F(2,82) = 9.15$, $p < .001$, $\eta_p^2 = .18$, with follow-up Tukey HSD comparisons showing significantly faster RTs for iconic images than both exemplars and averages (both *ps* < .05). There was not a significant difference in RTs between exemplars and averages ($p < .05$). For identities without iconic images, a related means *t*-test showed that RTs for average images were significantly slower than for exemplars $t(41) = 2.69$, $p = .01$, $d = .415$. These results show that iconic images were responded to faster

than averages, and the higher likeness ratings given to exemplars than averages translated into significantly faster RTs for the identities who did not have iconic images.

The results of this final experiment show that the perception of likeness is tied to a mental representation of a person insofar as there is a correlation between the likeness rating given to images and the speed at which they are recognised. In this experiment, we established three identities for whom there exists an iconic image. For these identities, the iconic images were given significantly higher likeness ratings than other, non-iconic images or an average image. The results also show that for identities without an iconic image, the average image is given a lower likeness rating than exemplars. The speeded name verification task showed faster RTs for iconic images than other exemplars and averages, and for some identities, slower RTs for averages than exemplars.

This exploration of likeness does seem to offer a route in to studying the central problem of familiar face representation. This experiment does not support the proposal that average faces are good candidate representations for recognition, in contrast to earlier proposals using different techniques (e.g. Burton et al, 2005). However, for familiar faces with iconic images, these do seem to have some privileged access to recognition, rendering them candidates for representations of those particular individuals. Of course, this leaves open a number of problems. Iconic images seem to be a property of only a rather restricted class of known individuals. It seems rather unlikely that our representations of personally familiar faces (family members etc) would be tied to particular images. Furthermore, as discussed above, only a rather restricted number of famous faces can be well-represented by particular iconic photos. Nevertheless, this finding does highlight the potential importance of 'picture-based' representations. Such representations are sometimes held to be at the root of unfamiliar face processing, but discarded for a more abstract coding as a face becomes familiar (Hancock, Bruce & Burton, 2000). In contrast, the findings reported here suggest a lasting influence of particular images for some familiar faces.

5. General Discussion

In three experiments, we have shown that the perception of 'what makes a good likeness' is tied to observers' representations of a person, and that these representations differ across observers based on their own experience. The results of Experiments 1 and 2 show that the more familiar we are with someone, the more we judge images of them to be a good likeness, and the faster we respond to images of them in a speeded name verification task. Across all three experiments, we have shown that higher likeness ratings for specific images also correlate with faster RTs for those images. These results suggest that observers are making coherent likeness judgements which do map on to the speed with which they recognise people, as well as their self-reported familiarity with those people. The results of Experiment 1 show that when observers rate images for likeness twice, they are quite consistent, and that 64% of the variance in likeness judgements can be explained by observers' private variance across images, with 36% of the variance shared across participants. This suggests that what constitutes a good likeness is to a greater extent particular to individual observers rather than being a generalised concept across people.

Experiment 2 provides further evidence for the idiosyncrasy of likeness judgements by tying judgements to specific instances of each celebrity that participants either had or had not previously seen by using images from films. The number of films seen for each celebrity correlated with familiarity with that celebrity. Moreover, images from films that participants had seen were rated as a higher likeness than images from films they had not seen.

In Experiment 3, we questioned what form our representations of familiar people take. We showed that for identities for whom an iconic image exists, that image is rated as a better likeness than other images. It has previously been suggested that the variability between different instances of familiar people is condensed into a stable average of those instances, and that this is how we represent familiar people (Burton et al., 2005; Jenkins & Burton, 2011). We therefore hypothesised that an average image may be rated as a higher likeness than other images. Instead, the results of Experiment 3 showed that average images are in fact rated as a poorer likeness than these exemplars. Moreover, average images were not responded to faster in a speeded name verification task, suggesting that they are not easier to recognise than other images of an identity. Our average images comprised ten instances of each person, taken from an internet search. The first two experiments in this paper show that idiosyncratic exposure to images plays a role in the perception of likeness, and so we cannot rule out that an average of previously seen images, or an average giving higher weighting to

recently-seen images would be rated as a good likeness. Of course, this experiment used a relatively small number of identities, and further experimentation will be necessary to assess the generality of this result.

We have used celebrity images as the familiar faces in the studies presented here. Naturally, this does not imply that the notion of familiarity is binary. For example, Carbon (2008) suggests that familiarity follows a three-tier hierarchy: personally familiar, famous and unfamiliar. Of course, familiarity is much more nuanced than this, and is a multi-faceted phenomenon which has been surprisingly little-studied in the literature on face recognition. It is important therefore to acknowledge that our manipulation of familiarity is very coarse and that more subtle variation is likely to be important – especially in personal life where familiarity is associated with other social roles and relationships. However, the results do make a start by demonstrating that familiarity is an important contributor to the notion of likeness.

Our results further suggest that the more familiar we are with someone, the more robust a representation of them we have, allowing us to easily incorporate new images of that person and judging each to be a good likeness. This ties in with previous work showing more consistent social judgements for different images of familiar compared to unfamiliar faces (Jenkins et al., 2011). There is a growing body of research which suggests that exposure to variability in a face helps us to learn what a new person looks like (Burton, Kramer, Ritchie, & Jenkins, 2016; Dowsett, Sandford, & Burton, 2016; Murphy, Ipser Gaigg, & Cook, 2015). In fact, we have recently shown that learning a new person from highly variable face images is more efficient than learning a person from the same number of images which show a smaller range of variability. Our results showed that learning new people from images varying in lighting, age, and hair styles produced benefits on a number of subsequent recognition tasks compared with learning from images which simply showed changes in pose and expression (Ritchie & Burton, 2017). The results reported in this paper support the converging evidence that exposure to variability within images of the same person allows us to recognise them in multiple novel images, because here, we have shown that the more familiar we are with someone, the more we will consider multiple varied images of them to be a good likeness. Repeated varied exposure to a person creates familiarity, as well as a tolerance for different images of that person. Heeding this intrinsic link between face

learning, familiarity, and the perception of likeness is key to furthering our understanding of the differences between familiar and unfamiliar face processing.

The effect of familiarity on likeness ratings is of particular relevance when considering the process of applying for a passport. Passport-issuing offices around the world require the applicant or someone familiar with them to verify the likeness of the passport image (Australian Passport Office, 2012; Passport Canada, 2013; HM Passport Office, 2014). Our results showing the correlation between familiarity and mean likeness rating for a range of images suggest that someone who knows the passport applicant well is likely to be more tolerant of an image which others may not consider to be a good likeness. We could therefore suggest, based on these results, that familiar people should not be asked to verify passport images. Recent research has shown that individuals asked to choose their own ID images do not pick the optimal photos. Instead, observers who have been briefly familiarised with the individual choose photos which are subsequently best matched by unfamiliar viewers. (White et al., 2016). It has been shown that a number of professional groups, such as police officers (Burton, Wilson, Cowan & Bruce, 1999), and passport officers (White, Kemp, Jenkins, Matheson & Burton, 2014; Writh & Carbon, 2017) perform surprisingly poorly at unfamiliar face matching, and show no association between experience and accuracy. We therefore suggest that images verified as a 'good likeness' by familiar viewers may be difficult for unfamiliar observers, such as passport or police officers, to recognise.

These experiments provide the first systematic investigation of the perception of likeness, and the first evidence showing that the representations we form of familiar people are idiosyncratic and tied to our individual experiences of these people.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.323262 to A. Mike Burton.

References

Allen, H., Brady, N., & Tredoux, C. (2009). Perception of 'best likeness' to highly familiar faces of self and friend. *Perception, 38,* 1821–1830.

Australian Passport Office (2012). Photograph guidelines. Retrieved from https://www.passports.gov.au/web/requirements/photos.aspx

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, *21*(3), 372-374.

Benson, P.J., & Perrett, D. I. (1991). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology, 3*(1), 105-135.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207–218.

Bruce, V., Ness, H., Hancock, P. J. B., Newman, C., & Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology, 87*(5), 894-902.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256-284.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40,* 202-223.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: evidence from security surveillance. Psychological Science, *10*(3), 243-248.

Carbon, C.-C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception, 37,* 801-806.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception, 31,* 985–994.

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11*(7), 857–869.

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *The Quarterly Journal of Experimental Psychology, 69*(1), 1-10.

Frowd, C. D., White, D., Kemp, R. I., Jenkins, R., Nawaz, K., & Herold, K. (2014). Constructing faces from memory: the impact of image likeness and prototypical representations. *Journal of Forensic Practice, 16*(4), 243-256.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337.

Hay, D. C., Young, A. W., & Ellis, A. W. (1991). Routes through the face recognition system. *The Quarterly Journal of Experimental Psychology A, 43*(4), 761-791.

HM Passport Office (2014). Countersigning passport application and photos. Retrieved from: https://www.gov.uk/countersigning-passport-applications

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgements of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 199-209.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671–1683.

Jenkins, R., White, D., van Montfort, X. & Burton, A.M. (2011). Variability in photos of the same face. *Cognition, 121,* 313-323.

Kramer, R. S. S., Gottwald, V. M., Dixon, T. A., & Ward, R. (2012). Different cues of personality and health from the face and gait of women. *Evolutionary Psychology, 10*(2), 271-295.

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from sets of images of the same person: A route to face learning. *Journal of Vision, 15*(4)1, 1-9.

Laurence, S., & Mondloch, C. J. (2016). That's my teacher! Children's ability to recognize personally familiar and unfamiliar faces improves with age. *Journal of Experimental Child Psychology, 143,* 123-138.

Lee, J. L., & Perrett, D. I. (2000). Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments. *Perception, 29,* 1291–1312.

Leikas, S., Verkasalo, M., & Lönnqvist, J. E. (2013). Posing personality: Is it possible to enact the Big Five traits in photographs? *Journal of Research in Personality, 47,* 15-21.

Monin, B., & Oppenheimer, D. M. (2005). Correlated averages vs. averaged correlations: Demonstrating the warm glow heuristic beyond aggregation. *Social Cognition, 23*(3), 257-278.

Murphy, J., Ipser, A., Gaigg, S., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577–581.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, USA, 105,* 11087-11092.

Passport Canada (2013). Photos: rules for Canadian passport photos. Retrieved from: http://www.ppt.gc.ca/info/photos.aspx

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19*(4), 473-497.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology.* Doi: 10.1080/17470218.2015.1136656

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M. White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141,* 161-169.

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2015). Face averages enhance user recognition for smartphone security. *PLoS One, 10,* e0119460.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, M. D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127,* 105-118.

Tidemann, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE, 21*(5), 42-50.

White, D., Burton, A. L., & Kemp, R. I. (2016). Not looking yourself: The cost of self-selecting photographs for identity verification. *British Journal of Psychology.* Advance online publication.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS One, 9,* e103510.

Wirth, B. E., & Carbon, C.-C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied, 23*(2), 138-157.